Session 2

Bayesian Methodology for Disclosure Limitation and Statistical Analysis of Large Government Surveys

Bayesian Methodology for Disclosure Limitation and Statistical Analysis of Large Government Surveys

Discussant: Ramesh Dandekar, Energy Information Administration, U. S. DOE

Researchers: Roderick J. Little and Trivellore Raghunathan, University of Michigan

Background

Synthetic micro data has been used extensively to study the behavior of complex computer models for a long time. In recent years, there has been an increased realization that synthetic micro data could also be used for a dissemination of statistical information in place of real data containing sensitive records collected by federal agencies. Because of relatively low disclosure potential and the ability to recreate most of the statistical properties of the original data, synthetic micro data offers some advantage over other methods of micro data protection. It has also been known for a while that synthetic data offers an economical choice to the on-site data research centers operated by federal statistical agencies in dissemination of public use information. Ideally, potential researchers could use synthetic data from their own work site for initial hypothesis testing/model development, without concern for data confidentiality. The researchers will need to use the data center facility only to run their final refined model/setup on the original data. Such a strategy has the potential to reduce the on-site operating cost for data centers.

The characteristics of micro data disseminated by federal statistical agencies vary considerably. As a result, it is unlikely that one synthetic micro data generation method will work well on all different micro data types. This necessitates that statistical agencies conduct a broad-based research on multiple fronts to generate synthetic micro data. The two separate papers in this session offer unique application areas.

The paper by Raghunathan, Reiter and Rubin, "Multiple Imputation for Statistical Disclosure Limitation", demonstrates the procedure to generate synthetic micro data by using multiple imputation framework proposed by Rubin in 1993. The proposed procedure uses a parametric and non parametric approach to generate synthetic data. The inference based on this technique requires that some adjustments be made to point and variance estimates prior to their use. The paper demonstrates that the inferences derived from the synthetic data are similar to those derived using actual data.

The Paper by Little and Liu, "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Micro Data", on the other hand, generates synthetic micro data by selective multiple imputation of categorical key variables and continuous non-key variables. The method offers a potential balance between data quality and statistical disclosure control by mixing select non-sensitive cases with sensitive cases.

Specific Comments

Both methods for synthetic micro data generation offer viable options by using a Bayesian framework. There are many potential applications for these two methods. However, the application potential for these two methods could be increased considerably by extending the scope of current research work to do the following:

1) Develop alternate methods/procedures to reduce current dependence on the model based imputation procedure. Developing the most appropriate global model to capture multi-variate statistical characteristics of any given data is always a time consuming process. It is also possible that the synthetic data end user might want to use the data to develop his/her own statistical model to represent original data. In such a situation, it might not be a good strategy to generate model-based synthetic data.

2) Derive new methods/procedures that will keep an optimum balance between the synthetic micro data quality and related tabular data quality along with adequate disclosure protection for both. It is a common practice to perform a preliminary statistical analysis of raw micro data by exploring associated tabular structure of the micro data. Conclusions derived from the tabular data analysis are commonly used in analytical studies and policy papers. Such a practice necessitates adequate precautions to retain statistical characteristics associated with original tabular structure to the extent possible.

3) Look at the feasibility of using the Latin Hypercube Sampling (LHS) method in combination with a restricted pairing algorithm by Iman and Conover to induce a desired rank correlation matrix on synthetic micro data within a framework supported by a Bayesian method. The LHS method is model independent and has been used successfully to generate synthetic micro data since late seventies. By using the empirical cumulative distribution function of the real data, the LHS method provides non-parametric approach to generate synthetic micro data. For many applications the LHS-based synthetic data generation method could offer the most practical approach that balances data quality and minimal resources required to generate synthetic micro data.

4) Look at the feasibility of performing backward calibration of micro data based on the outcome from the Controlled Tabular Adjustments (CTA) to protect related tabular data (Dandekar/Cox 2002, Dandekar 2003). Such a strategy allows one to one correspondence between synthetic micro data and synthetic tabular data.

References

Dandekar, Ramesh A. (1993), "Performance Improvement of Restricted Pairing Algorithm for Latin Hypercube Sampling", ASA Summer conference (unpublished).

Dandekar Ramesh A. and Cox Lawrence H. (2002), Synthetic Tabular Data: An Alternative to Complementary Cell Suppression, manuscript available from ramesh.dandekar@eia.doe.gov.

Dandekar R. A., Cohen M., and Kirkendall, N. (2002a), Sensitive Micro Data Protection using Latin Hypercube Sampling Technique. In J. Domingo-Ferrer, ed., Inference Control in Statistical Databases, 117-125., Berlin:Springer-Verlag

Dandekar, R.A (2003), Cost Effective Implementation of Synthetic Tabulation (a.k.a. Controlled Tabular Adjustments) in Legacy and New Statistical Data Publication Systems, working paper 40, UNECE Work session on statistical data confidentiality (Luxembourg, 7-9 April 2003)

Iman R.L. and Conover W. J. (1982), "A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables", Commun. Stat., B11(3): pp. 311-334.

McKay M.D., Conover W. J. and Beckman, R. J. (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code", Technometrics 21(2): pp. 239-245.

Discussion of

Multiple Imputation for Statistical Disclosure Limitation by T. E. Raghunathan, J. P. Reiter, and D. B. Rubin

Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata by R. J. A. Little and F. Liu

William E Winkler
U.S. Census Bureau

1. Introduction

Statistical agencies that provide public-use microdata must contend with the conflicting goals of producing data that satisfy one or more analytic needs of a group of users and preserving the confidentiality of data records associated with entities such as individuals or companies. It is the view of this discussant (e.g., Winkler 1997) that analytic needs should be met by building models of the public microdata. The models should be described in terms of user-specified requirements for analyses. The documentation should describe the limitations the microdata for the specified analytic purposes and other purposes to which the microdata might be put. If the analytic needs of the microdata have been justified, then the confidentiality of the microdata should have be described.

The outline of this discussion is as follows. In second section, I provide background on a number of existing methods and their analytic limitations. In the third section, I discuss the general framework of Raghunathan et al. (2003) for providing synthetic microdata under models that meet analytic needs and the framework of Little and Liu (2003) for providing partially synthetic data that also meets analytic needs and does not require the amount of modeling as the more general framework. The final section consists of concluding remarks.

2. Background

A variety of methods have been developed and used for masking a data file. The methods have the intent of altering the data in a manner that allows some analyses to be done that correspond to what could be done on the original, confidential microdata and of making re-identification more difficult. After masking, the resultant microdata are disseminated to users who presumably wish to perform analyses that could not be performed by using published tables alone.

These masking methods include swapping (Dalenius and Reiss 1982), rank swapping (Moore 1996), micro-aggregation (e.g., Domingo et al. 2002), k-similarity (Samarati and Sweeney 1998) that includes global recoding and local suppression, variants of additive noise (Kim 1986, 1990, Fuller 1993), and synthetic microdata (Rubin 1993, Fienberg 1997). All of the original and succeeding

authors who have considered swapping, rank swapping, and micro-aggregation have been able to point out serious difficulties with providing for even basic analytic needs. If the swapping, rank swapping, and micro-aggregation are over relatively small and homogenous groups, then simple analytic needs may not be seriously compromised but re-identification can be straightforward (Winkler 2002). The method of Winkler (2002) for micro-aggregation it can be easily extended to swapping and rank swapping. Although k-similarity is guaranteed to provide confidentiality because at least k records with have the same identifying information, it has, so far, only been rigorously shown to provide analytic needs in very simple situations (Iyengar 2002). Sampling, as a simple alternative, neither assures that simple analytic needs are met nor assures that all records cannot be re-identified. Typically, sampling is not designed to satisfy a number of analytic constraints (particularly on a set of subdomains). With typical sampling designs, records in the sample can be population uniques and relatively straightforward to re-identify.

The only two methods that place primary emphasis on analytic properties of the masked microdata are the additive noise ideas of Kim (1986, 1990) and synthetic data methods (Rubin 1993, Fienberg 1997). A valid criticism of additive noise has been that it is only generally suitable for public-use microdata that is used in regression-type analyses. Another criticism has been that special software is needed for analyzing additive-noise microdata. High quality software (Yancey et al. 2002) is now available for correct analysis. The software even supports analyses on arbitrary subdomains according to the original ideas introduced by Kim (1990). At present, producing synthetic data according to models that consider user-specified analytic needs are the most promising approach. Criticisms of the approach deal with the inability of groups, particularly in statistical agencies, to develop models of their data and create software. A simplistic method for automatically creating models of the data using Bayesian networks was introduced by Thibaudeau and Winkler (2002). The standard methods for creating models for multiple imputation should still produce much higher quality analytic properties.

3. The Papers

This section summarizes and comments on the papers of Raghunathan, Reiter and Rubin (2003) and Little and Liu (2003).

3.1. Raghunathan, Reiter, and Rubin

The paper of Raghunathan et al. (2003) provides an important theoretical foundation for producing synthetic microdata satisfying analytic constraints. Three examples give insight and provide further practical advice. Other examples have been given by Reiter (2002, 2003). Further, software (Raghunathan et al. 1998) can facilitate producing microdata in a manner that is consistent with ideas introduced by Kennickell (1997) and Abowd and Woodcock (2002).

Fienberg (1997) raised the following issue. If sufficient analytic constraints are placed on the synthetic microdata, then some of the synthetic microdata records may be very close to actual population records. This has the possibility of allowing re-identification. In the Raghunathan et al. (2003) framework, arbitrary statistics q_M and T_M representing multiple imputation means and variances are considered. If a sufficiently large number of copies of the population P_i, $i \leq M$, are released and the models are sufficiently detailed to allow reasonable analyses on a moderate number of statistics q, when will it be possible that a moderate number of the original, confidential microdata

records may be approximated with reasonable accuracy? Raghunathan et al. note that the approximate Bayesian bootstrap, while not as sensitive to model assumptions, can potentially lead to more re-identification. The parametric modeling, on the other hand, is more subject to model specification error as has been noted by Reiter (2002) in addition to Raghunathan et al..

3.2. Little and Liu

The paper of Little and Liu (2003) provides a practical framework for producing partially synthetic data that should be more straightforward to implement than purely synthetic data. Their paper also provides a useful and practical guide about how to do re-identification in straightforward situations.

I summarize their method. They assume that the original data consist of both continuous and discrete variables. They assume that outside individuals have a database that contains the discrete variables. Their method "masks" the discrete variables in a manner that does not change the continuous variables. They mask by choosing neighborhoods of variables using continuous variables only. Discrete data among "at risk" records or merely in a sample of records within the neighborhoods can be swapped. There is no requirement that the neighborhoods are disjoint. Their initial empirical results are promising. They demonstrate that the information loss due to the masking procedure is modest but still non-trivial. Using discrete data only, they provide re-identification risk metrics that are conservative and realistic.

If both continuous and discrete data are used for re-identification, is it possible to re-identify? Little and Liu might compare their information-loss/re-identification-risk framework to the R-U confidentiality map framework introduced by Duncan et al. 2001 (see also Trottini and Fienberg, 2002).

4. Concluding Remarks

The concluding remarks are two recommendations. The first recommendation is that all releases of public-use microdata should discuss and justify the analytic usefulness of the data. This should include what analyses on the original, confidential microdata can be reproduced on the masked, public-use microdata. The second recommendation is that the microdata confidentiality community should continue serious investigation of synthetic microdata, particularly with the information-loss/disclosure-risk framework given by both sets of authors. An alternative method for producing synthetic microdata using Latin Hypecubes is given by Dandekar et al. (2002).

References

Abowd, J. M. and Woodcock, S. D. (2002), "Disclosure Limitation in Longitudinal Linked Data," in (P. Doyle et al., eds.) *Confidentiality, Disclosure, and Data Access,* North Holland: Amsterdam.

Dalenius, T. and Reiss, S.P. (1982), "Data-swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inference*, **6**, 73-85.

Dandekar, R. A., Domingo-Ferrer, J. and Sebe, F. (2002), "LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York.

Dandekar, R., Cohen, M. and Kirkendal, N. (2002), "Sensitive Microdata Protection Using Latin Hypercube Sampling Technique," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002), "Practical Data-Oriented Microaggregation for

Statistical Disclosure Control," *IEEE Transactions on Knowledge and Data Engineering*, **14** (1), 189-201.
Duncan, G. T., Keller-McNulty, S. A.,and Stokes, S. L. (2001), "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map," Los Alamos National Laboratory Technical Report LA-UR-01-6428.
Fienberg, S. E. (1997), "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, commissioned by Committee on National Statistics of the National Academy of Sciences.
Fienberg, S. E., Makov, E. U. and Sanil, A. P., (1997), "A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data," *Journal of Official Statistics*, **14**, 75-89.
Fienberg, S. E., Makov, E. U. and Steel, R. J. (1998), "Disclosure Limitation using Perturbation and Related Methods for Categorical Data," *Journal of Official Statistics*, **14**, 485-502.
Fuller, W. A. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, **9,** 383-406.
Iyengar, V. (2002), "Transforming Data to Satisfy Privacy Constraints," Association of Computing Machinery, Special Interest Group on Knowledge Discovery and Datamining '02.
Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 303-308.
Kim, J. J. (1990), "Subdomain Estimation for the Masked Data," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 456-461.
Kim, J. J. and Winkler, W. E. (1995), "Masking Microdata Files,"American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 114-119.
Little, R. J. A. (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics*, **9,** 407-426.
Little, R. J. A. and Liu, F. (2002), "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, to appear.
Little, R. J. A. and Liu, F. (2003), "Comparison of SMIKe with Data-Swapping and PRAM for Statistical Disclosure Control of Simulated Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, to appear.
Moore, R. (1995), "Controlled Data Swapping Techniques For Masking Public Use Data Sets," U.S. Bureau of the Census, Statistical Research Division Report rr96/04, (available at http://www.census.gov/srd/www/byyear.html).
Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., and Sollenberger, P. (1998), "A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models," Survey Research Center, University of Michigan.
Raghunathan, T.E., Reiter, J. P. and Rubin, D.R. (2003), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, **19**, 1-16.
Reiter, J.P. (2002), "Satisfying Disclosure Restrictions with Synthetic Data Sets," *Journal of Official Statistics*, **18**, 531-543.
Reiter, J.P. (2003), "Methods of Inference for Partially Synthetic, Public Use Data Sets," *Journal of Official Statistics*, to appear.
Rubin, D. B. (1993), "Satisfying Confidentiality Constraints through the Use of Synthetic Multiply-imputed Microdata,"*Journal of Official Statistics*, **91**, 461-468.
Samarati, P. (2001), "Protecting Respondents' Identity in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering*, **13** (6), 1010-1027.
Samarati, P. and Sweeney, L. (1998), "Protecting Privacy when Disclosing Information: k-anonymity and its Enforcement through Generalization and Cell Suppression," Technical Report, SRI International.
Sweeney, L. (2002), "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 571-588.
Thibaudeau, Y. and Winkler, W.E. (2002), "Bayesian Networks Representations, Generalized Imputation, and Synthetic Microdata satisfying Analytic Restraints," Statistical Research Division report RRS 2002/09at http://www.census.gov/srd/www/byyear.html.
Trottini, M. and Fienberg, S. E. (2002), "Modelling User Uncertainty for Disclosure Risk and Data Utility," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 511-528.
Winkler, W. E. (1997), "Views on the Production and Use of Confidential Microdata," Statistical Research

Division report RR 97/01 at http://www.census.gov/srd/www/byyear.html.
Winkler, W. E. (2002), "Single Ranking Micro-aggregation and Re-identification," Statistical Research Division report RRS 2002/08 at http://www.census.gov/srd/www/byyear.html.
Yancey, W.E., Winkler, W.E., and Creecy, R. H. (2002) "Disclosure Risk Assessment in Perturbative Microdata Protection," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, Springer: New York (also report RRS 2002/01 at http://www.census.gov/srd/www/byyear.html)